WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

Sequential Screening in Semiconductor Manufacturing, I: Exploiting Lot-to-Lot Variability

Jihong Ou

and Lawrence M. Wein

WP# 3451-92-MSA July, 1992

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

Sequential Screening in Semiconductor Manufacturing, I: Exploiting Lot-to-Lot Variability

Jihong Ou

and Lawrence M. Wein

WP# 3451-92-MSA July, 1992

SEQUENTIAL SCREENING IN SEMICONDUCTOR MANUFACTURING, I: EXPLOITING LOT-TO-LOT VARIABILITY

Jihong Ou

Operations Research Center, M.I.T.

and

Lawrence M. Wein

Sloan School of Management, M.I.T.

Abstract

We address a problem of simultaneous quality and quantity control motivated by semiconductor manufacturing. After wafers are fabricated, they are probed, or electrically tested, and in some cases the probing facility is the bottleneck for the entire IC manufacturing process. Under this assumption, we consider the problem of choosing the optimal start rate of lots of wafers into the fabrication facility and the optimal screening policy in front of the probing facility to maximize the expected profit, which is the revenue from good chips minus the variable fabrication and probing costs. The screening policy decides which wafers to discard and which wafers to probe. These decisions are subject to capacity constraints at both the wafer fabrication and probing facilities. An empirical Bayes approach is employed: the number of bad chips on a wafer is assumed to be a gamma random variable, where the scale parameter is unknown and varies from lot to lot according to another gamma distribution. We fit the yield model to industrial data and test the optimal policy on this data.

July 1992

SEQUENTIAL SCREENING IN SEMICONDUCTOR MANUFACTURING, I: EXPLOITING LOT-TO-LOT VARIABILITY

Jihong Ou

Operations Research Center, M.I.T.

and

Lawrence M. Wein

Sloan School of Management, M.I.T.

This paper and its companion (Longtin et al. 1992) address a particular quality management issue in semiconductor manufacturing. The production of integrated circuits consists of four main stages: wafer fabrication, probing, packaging and final testing, and we will focus on the interrelationship between wafer fabrication and probing. In wafer fabrication, disc-like wafers that contain hundreds of integrated circuits, or *chips*, are produced (in batches, or *lots*, of usually 20 to 50 wafers) by a very long and complex procedure involving hundreds of operations. After fabrication is completed, each chip on a wafer is *probed*, or electrically tested, to distinguish between defective chips and good chips. Each wafer is then separated into its respective chips, and nondefective chips are covered in a protective plastic during packaging. In final testing, the chips are functionally tested under a variety of environmental conditions before being shipped to customers.

Problem Description. To motivate our model formulation, the process economics and the key material flow issues need to to be briefly described. Building a wafer fabrication facility, or *fab*, costs hundreds of millions of dollars, and consequently, fab managers are very concerned with maintaining high utilization of the bottleneck equipment, and one of the biggest operational decisions for the fab manager is to determine the *start rate* of wafers into the fab. Because of the huge amount of statistical variability in the fab (due primarily to random yield, rework, and tool failures; see Chen et al. 1988 for details), pushing the start rate beyond a certain level, which we call the fab's *effective*

capacity, will result in unacceptably high levels of work-in-process inventory and long lead times.

Despite the well-documented congestion that occurs in wafer fabrication, we have visited several facilities where the probing (we will often use the more generic term *testing* rather than probing) facility, not the wafer fab, is the bottleneck that determines the production capacity of finished goods. There are several reasons for this phenomenon: the testing equipment is very expensive (machines can cost three to four million dollars) and the testing procedure is a very time consuming and labor intensive process. Furthermore, testing capacity is sometimes labor constrained because companies are either unable or unwilling to hire and train full-time employees in the face of uncertain future demand.

Although relative costs and revenues depend greatly on the type of market (e.g., commodity or custom chips) and other factors, the variable testing cost per wafer is typically only several percent of total variable production cost, and the revenue from a wafer of nondefective chips is roughly ten times the variable production cost per wafer. Also, the *yield* in wafer fabrication, which is the fraction of chips that are good, can be very low and erratic. Since many facilities are capacity constrained rather than market constrained, they can sell anything they make, and any increase in yield leads directly to an increase in profit. Consequently, yield dominates the economics of the process and is the primary concern of fab managers.

Semiconductor manufacturers typically use an *exhaustive* testing policy; that is, every chip of every wafer is tested and is deemed defective or nondefective. Indeed, the thought of simply discarding a completed chip before it undergoes testing (unless it represents "leftovers" from a custom order that has already been filled) goes very much against the grain of mainstream industry thinking. In contrast, our paper and Longtin et al. are based on the following simple premise that has also been put forth in Goldratt and Cox (1984): *profitability can be increased by preventing bottleneck equipment from working on products that are already defective.* In particular, if testing is the bottleneck operation under an exhaustive testing policy, then semiconductor manufacturers can increase their

profits by simultaneously (1) employing a sequential screening procedure that adaptively discards, rather than tests, portions of wafers (or entire wafers or even entire lots) that are thought to have a sufficiently low proportion of nondefective chips, and (2) increasing the start rate of wafers. Of course, if the rate of wafer starts is increased, so is the congestion in the fab and the production costs, and these two factors need to be taken into account.

To test this premise, we consider the following problem of *simultaneous quality and quantity control*: determine the start rate of lots of wafers into the fab and find a sequential screening policy for the testing facility to maximize the expected long run average revenue from nondefective chips minus the variable fabrication and testing costs of wafers. The two controls are subject to constraints on the average effective capacity of both the fab and the testing station; we assume that the testing capacity constraint is more restrictive than the fab constraint when an exhaustive testing policy is in use. To minimize confusion, a screening procedure in isolation will be referred to as a *policy* and a screening policy coupled with a start rate will be referred to as a *strategy*.

In practice, the resulting increase in profit that an optimal strategy will achieve relative to the exhaustive testing strategy commonly used in industry (that is, an exhaustive testing policy with a start rate that keeps the testing facility working at its effective capacity) depends greatly on two factors that will be discussed below: (1) the relative congestion levels of the fab and the testing facility under the exhaustive testing policy, and (2) the nature of the yield variability. Indeed, if the fab *was* more highly congested than the testing facility under an exhaustive testing policy, then this policy might be optimal, and hence sequential screening would be of no value. However, testing is also performed after various key operations in the fab, and often (for example, see the simulation models of Atherton and Dayhoff 1985, Glassey and Resende 1988 and Wein 1988) the bottleneck workstation in the fab is the photolithography workstation, to which wafers make many (up to twenty) visits during their processing. Thus, the framework presented here can also be used to perform sequential screening at key tests in the fab. That is, the start rate of wafers can be increased and undesirable wafers or chips can be discarded at

in-fab tests so that the bottleneck equipment works on higher quality chips. However, this procedure may not be as effective in the fab as it is at probe, because type I and type II errors are apt to be more prevalent in the fab. In particular, in-fab testing is often visual and is not as discerning as electrical testing, and a chip that is correctly found to be nondefective at an in-fab test may become defective before its next visit to the bottleneck workstation. We will hereafter assume that testing is the bottleneck operation, and more specifically, our numerical studies here and in Longtin et al. assume that the fab is at 90% of its effective capacity when the exhaustive testing strategy is employed.

Yield Modeling. We now discuss the nature of the yield variability. Low yield in wafer fabrication occurs for a variety of reasons, including short product life cycles, particulate contamination (see Osburn et al. 1988), misalignment of operations, and chemical imbalances. Also, defective chips are difficult to detect visually, and the industry relies heavily on the probing machines. Intuitively, sequential screening will only be effective if *dependencies and/or nonuniformities in yield can be identified and exploited.* After all, if every chip processed by the fab had the same probability of being defective, independently of all other chips, then sequential screening would be fruitless. However, several types of dependencies do exist and, indeed, one of our primary goals in this pair of papers is to analyze industrial data and determine which dependencies are most prevalent and easiest to exploit.

Recall that chips are produced on wafers, and wafers travel through the fab in lots. Dependencies may be present at all three levels (lots, wafers, chips), including

(1) *dependence across consecutive lots*: the yield of consecutive lots may be positively correlated because of machines that go in and out of control, or batch operations, such as diffusion or oxidation, that simultaneously process multiple lots;

(2) *nonuniformity in chip type*: some chip types may be inherently easier to produce than others;

(3) *dependence of wafers within a lot*: positive serial correlation of wafer yields within a lot may be due to operations that simultaneously process one or more lots of wafers, or to wafer-by-wafer operations that incur a joint set-up for an entire lot;

(4) *dependence of neighboring chip locations on a wafer*: defective chips are often found in clusters (see Mallory et al. 1983 for empirical data), which may be due to processing or particulate contamination;

(5) *radial nonuniformity on a wafer*: handling and processing can cause a donut-shaped yield with more defective chips on the edge of the wafer and, to a lesser extent, in the center of the wafer (see Ferris-Prabhu et al. 1987 for empirical data); and

(6) *dependence of a chip location across wafers within a lot*: mask defects and batch operations can cause the yield of a chip location to be positively correlated across consecutive wafers.

Furthermore, sequential screening can be performed at all three levels: we can discard (i) entire lots of wafers based on the yield from previous lots, (ii) wafers in a lot based on the yield of previously tested wafers from the same lot, or (iii) chips on a wafer based on the yield of previously tested chips. We do not pursue screening of type (i) because dependency (1) is not very prevalent in wafer fabrication; since a wafer fab is far from a flow line operation, lots that are processed together in the same oven during a particular batch operation tend to go their separate ways and do not arrive together at the testing facility. Also, all the industrial data sets that we analyze contain lots of only one chip type, and hence nonuniformity (2) will not be addressed. However, this factor could be addressed in our framework by developing a different yield model for each type of chip.

Our two studies employ sequential screening of types (ii) and (iii) to exploit dependencies (3)-(6). The factors underlying dependency (3), coupled with the high degree of randomness in the production process, lead to a significant amount of *lot-to-lot variability in mean yield*. In this paper, sequential screening of type (ii) is employed to exploit lot-to-lot variability. Our yield model assumes that the number of defective chips on each

wafer in a given lot is an independent gamma random variable with shape parameter α and scale parameter β. An empirical Bayes approach is used, where the scale parameter β is unknown and varies from lot to lot; for each lot, the parameter β is chosen independently from a (different) gamma prior distribution. A sequential screening policy in this setting decides when to *discard the remaining wafers in a lot.*

Industrial data from Bohn (1991) is used to estimate the parameters for the yield model in this paper. Since the primitive empirical data in Bohn is the number of nondefective chips on each wafer, this data cannot be used to analyze the more detailed spatial and temporal dependencies described in (4)-(6). Longtin et al. analyze over 300 *wafer maps* (see the Appendix of that paper for some examples) from two wafer fabs, and model the chip yield by a Markov random field, which is a stochastic model that allows the probability of a chip being nondefective to depend on the resulting yield of the neighboring chips. A variety of sequential screening strategies of type (iii) are proposed that *discard individual chips on a wafer.* In summary, the present paper employs sequential screening at the wafer level to exploit lot-to-lot variability and Longtin et al. employ sequential screening at the chip level to exploit detailed spatial dependencies within a lot.

The two key aspects of yield modeling that we focus on, lot-to-lot variablity and spatial dependence on and across wafers, have received very little attention in the IC yield modeling literature. We know of no models capturing the former aspect and Flack (1985) appears to contain the only yield model that explicitly accounts for spatial dependence of chips on a wafer. Nearly all the existing yield literature (see Cunningham 1990 for a recent survey) calculates the proportion of nondefective chips on a wafer by considering the chip area and density of point defects on the wafer. These derivations lead to a two parameter distribution (the negative binomial distribution, which describes a Poisson random variable mixed with a gamma, appears to be the most effective) that can be fitted to the mean and variance of the empirical data for the number of nondefective chips per wafer. According to Cunningham, the goal of most of the chip yield modeling

research has been to predict costs and actual yields, and to determine the appropriate level of circuit integration. Albin and Friedman's (1989) work on acceptance sampling appears to be the first to employ a yield model in a quality control context; they use a two parameter distribution (the Neyman type A, which is a Poisson compounded Poisson) to model the number of defective chips on a wafer. Because they were interested in quality control issues rather than circuit design issues, they directly modeled the yield without resorting to the defect density and chip area, and we do the same in this pair of papers.

Summary of Results. The optimization problem addressed in this paper is essentially an optimal stopping problem embedded within a mathematical program, and the optimal solution is determined numerically by solving a series of parameterized optimal stopping problems. Since the optimal strategy is difficult to calculate, we also find the optimal fixed sample size strategy, where a fixed number of wafers from each lot is tested, after which the controller either discards or tests all the remaining wafers in a lot. Five of Bohn's industrial data sets are used to estimate the parameters of the yield model, and the two proposed policies are derived for all five data sets. For our parameter values, the maximum possible profit increase that an optimal strategy can achieve relative to the exhaustive strategy commonly used in industry is between 11.1% and 12%; the exact upper bound cannot be mentioned without revealing the true yield of Bohn's wafers. The fixed sample size strategy and the optimal strategy achieve a 2.2% and 2.5% average profit increase over the five data sets, respectively.

These two strategies are also tested on the actual data in a simulation study. By randomly shuffling the wafers in a lot, 100 lots of wafers are generated from each lot in the five data sets. If the yield model underestimates the average number of discarded wafers per lot in the simulation study, then the testing facility will be underutilized and a suboptimal strategy is obtained. If the yield model overestimates the average number of discarded wafers, then the testing facility will be overutilized, and an infeasible solution can result. Under the fixed sample size strategy, the model accurately predicts the average number of discarded wafers per lot, and an average profit increase of 1.2% is achieved.

Under the optimal strategy, the model underestimates the average number of discarded wafers by an average of 2.5% over the five data sets, which results in an average profit decrease of 0.7%.

In summary, the fixed sample size strategy may be preferable to the optimal strategy, since it is much easier to derive and to implement, it performs nearly as well on the analytical model, and appears to be more robust when faced with the actual data sets. We believe that the discrepancy between the theoretical results and the simulation results is due primarily to the assumption that all lots in the same data set have the same shape parameter α. Hence, a relaxation of this assumption is probably required to obtain a more accurate estimate of the average number of discarded chips per lot, which should lead to a more effective and reliable strategy. The profit increases reported here are relatively small and, in particular, are significantly smaller the the increases achieved by screening at the chip level in Longtin et al. However, readers should keep in mind that a 1% increase in revenue minus variable cost can represent millions of dollars annually. Also, since the fixed cost component is so large in this industry, a 1% improvement here would translate into a much larger percentage improvement in a company's reported profits.

The remainder of the paper is organized as follows. In Section 1, the yield model is described in detail, and the modeling assumptions are compared with the conclusions of Bohn's empirical study. The stochastic optimization problem is formulated in Section 2. The optimal fixed sample size screening strategy is found in Section 3, and the optimal sequential screening strategy is derived in Section 4. Numerical results are reported in Section 5. Concluding remarks on this paper and Longtin et al. can be found in Section 7 of the latter paper.

1. The Yield Model and Industrial Data

Our yield model assumes that the number of *defective* chips on each wafer in a given lot

is an independent gamma random variable with shape parameter α and scale parameter β. An empirical Bayes approach is used, where the shape parameter α is the same for all lots but the scale parameter β varies from lot to lot; for each lot, the value of the parameter β is chosen independently from a gamma prior distribution with known parameters a and b. The two gamma distributions form a *conjugate pair*: if the parameter β has a gamma (a, b) distribution prior to testing a wafer, and if x chips on the wafer are found to be defective, then β has a gamma $(a + \alpha, b + x)$ posterior distribution.

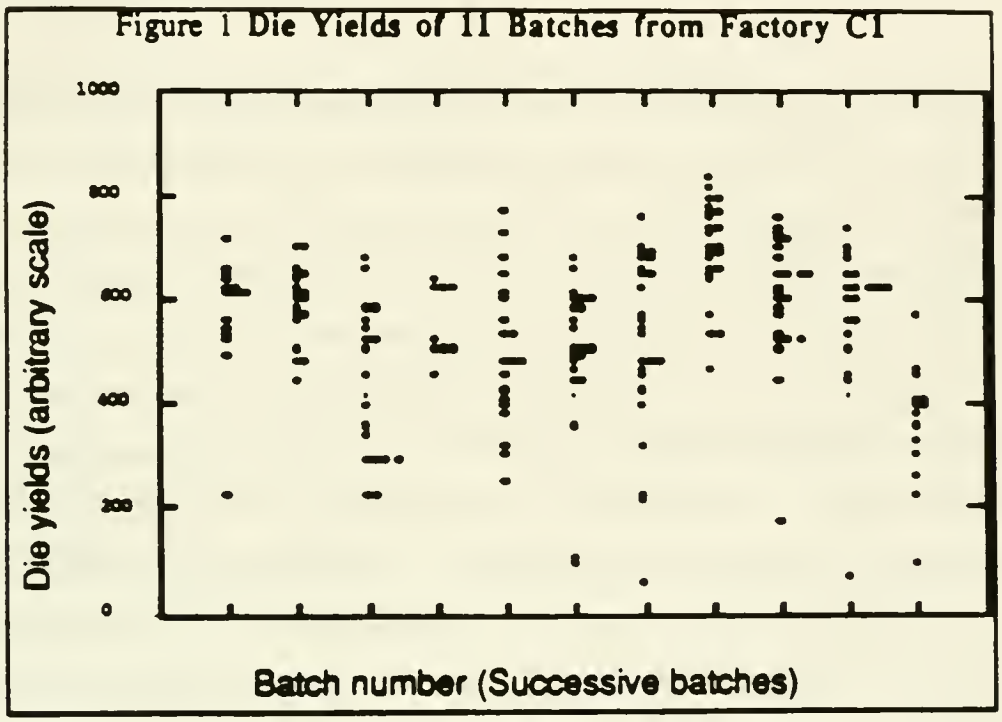

Figure 1. Figure 1 of Bohn.

In Section 5, maximum likelihood estimation is employed to estimate the parameters α, a and b from the industrial yield data collected by Bohn. He analyzed 11 different sets of data from five factories, and we analyze five of these data sets (sets C1,C1.5,C2,C2.5,C3), which are all from the same factory. Each data set consisted of about ten lots, or batches, of wafers of the same product that were completed during the same week. To disguise the actual yield, Bohn presented the raw data as the number

of good chips, or die, on each wafer of each lot. In Figure 1, we reproduce Figure 1 of Bohn, which displays a summary of data set C1. Each column in Figure 1 corresponds to a lot of wafers and each point represents the number of good chips on a particular wafer; hence, Figure 1 essentially contains 11 yield histograms, one for each lot. Bohn came to the following three conclusions concerning his 11 data sets: (i) the mean yield of each lot varies considerably from lot to lot (e.g., compare the last two lots); (ii) the within lot variability (i.e., the vertical spread of points in each column) is high; and (iii) there is a high variation between lots of within lot variability (e.g., compare the second and seventh lots).

The gamma-gamma model certainly captures the lot-to-lot variability in mean yield. However, plenty of other conjugate pairs would also capture this effect. In fact, before considering the gamma-gamma pair, we performed our entire analysis using the beta-binomial pair and the gamma-Poisson pair; the beta-binomial model, in particular, has intuitive appeal, since the number of bad chips per wafer is explicitly modeled as an integer between zero and the number of chips on a wafer. However, the binomial and Poisson assumptions significantly underestimate the within lot variability of chip yield. More specifically, we calculated the mean and variance of the number of good chips on each wafer of a given lot, and determined the variance-to-mean ratio for each lot in the five data sets. The average ratio over all 53 lots was 7.6 and the range was 1.8 to 30.3. In contrast, the corresponding variance-to-mean ratio under the binomial (Poisson, respectively) assumption is less than (equal to, respectively) one. Consequently, when the controls derived from these two yield models were tested on the actual data, too many wafers were discarded at the testing facility, which led to a significant reduction in overall profit. Although the gamma-gamma conjugate pair captures the substantial level of within lot variability, it is unable to capture the high variation between lots of within lot variability. Perhaps a gamma-gamma model in which both the shape and scale parameters are unknown would capture this effect; computational considerations prevented us from pursuing this avenue. In summary, the gamma-gamma model captures the effects in conclusion (i) and (ii), but does not capture the effect in conclusion (iii).

To further investigate the validity of our model, we test the derived policies on the actual data sets in Section 5.

Finally, readers may note that the number of wafers in a lot is not constant in Figure 1. This is due to the scrapping of entire wafers during fabrication. Hence, we also assume that each wafer in a lot has a certain probability of being scrapped during fabrication, so that the size of a lot exiting the fab is a binomial random variable.

2. Problem Formulation

In this section, we mathematically formulate the optimization problem described in the Introduction, and pictured in Figure 2. Each lot entering the fab contains L wafers and each wafer consists of M chips. Each wafer in a lot is scrapped during fabrication with probability q, and hence the lot size l of a wafer exiting the fab is a binomial random variable with parameters L and $1-q$. Since the exact number of wafers in a lot is known when the lot arrives to the testing facility, it is natural to use this information to develop an optimal screening policy. However, this would require us to derive a different optimal screening policy for every possible value of l, which makes the optimal solution much more difficult to compute and harder to implement in practice. Instead, we do not allow our screening policy to differ from lot to lot, except for the obvious constraint that no more than l wafers can be tested from a lot with l wafers. Since most fabs typically scrap about 5-10% of their wafers, this assumption should not lead to significant degradations in performance.

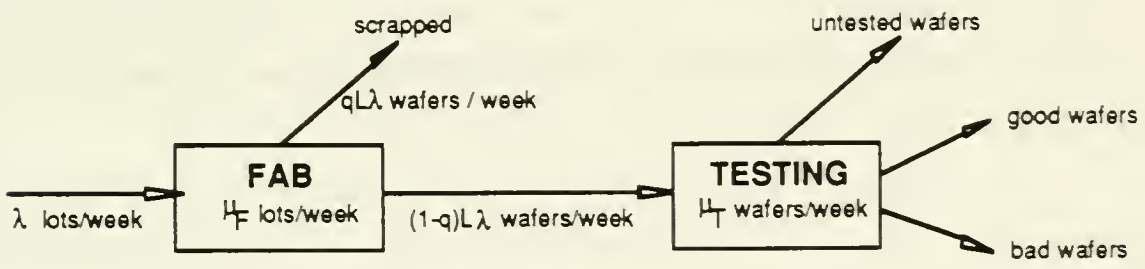

Figure 2. The semiconductor manufacturing facility.

For a typical lot exiting the fab, let x_n be the number of defective chips on the nth wafer, for $n = 1, ..., l$, and let $s_n = \sum_{k=1}^{n} x_k$ be the total number of defective chips on the first n wafers. Let the decision variable $u_n = 1$ if the nth wafer is to be tested and $u_n = 0$ if the nth wafer is to be discarded. A screening policy is defined by the vector $u = (u_1, u_2, ..., u_L)$, where $u_n = 0$ for $n > l$. For $n = 1, ..., l$, the decision μ_n depends on $(x_1, ..., x_{n-1})$ only through the sufficient statistic s_{n-1}, where $s_0 = 0$. The profit generated by wafer n is

$$g(x_n, u_n) = \begin{cases} 0 & \text{if } u_n = 0, \\ r(M - x_n) - c_T & \text{if } u_n = 1, \end{cases} \tag{1}$$

where r is the revenue received from a good chip and c_T is the variable testing cost per wafer. For a given policy u, the expected profit from testing one lot of wafers is

$$V(u) = E[\sum_{n=1}^{L} g(x_n, u_n)], \tag{2}$$

and the expected number of wafers tested per lot is

$$N(u) = E[\sum_{n=1}^{L} u_n], \tag{3}$$

where both expectations are over the random variables l, which is embedded in the definition of μ_n, and $(x_1, ..., x_l)$.

The problem of finding a screening policy u that maximizes (2) is an optimal stopping problem. Our problem of *simultaneous quality and quantity control* involves one more decision variable and two extra constraints. The decision variable is the lot start rate λ, which is the number of lots introduced into the fab per week. Let the fab's *effective capacity* be μ_F lots per week and let the testing facility's effective capacity be μ_T wafers per week. We assume that if the rate of work entering either of these facilities exceeds its effective capacity, then unacceptably high lead times and work in process inventory levels will be incurred. Hence, the two constraints are

$$\lambda \leq \mu_F \tag{4}$$

and

$$\lambda N(U) \leq \mu_T. \tag{5}$$

Our optimization problem is to choose the start rate λ and a screening policy u to maximize

$$\lambda(V(u) - c_F) \tag{6}$$

subject to constraints (4) and (5), where c_F is the variable fabrication cost per lot.

We conclude this section with some assumptions on the problem parameters. Before any testing is performed, the a priori expected number of defective chips on a wafer is

$$E[x_n] = \frac{\alpha b}{a-1}. \tag{7}$$

To ensure that this quantity is positive, we need to assume that $a > 1$, which holds for the parameter estimates obtained from Bohn's data in Section 5. If we denote the exhaustive testing policy by u^E, then

$$V(u^E) = (1-q)L[r(M - \frac{\alpha b}{a-1}) - c_T], \tag{8}$$

and

$$N(u^E) = (1-q)L. \tag{9}$$

We assume

$$\mu_F > \frac{\mu_T}{(1-q)L}, \tag{10}$$

so that the testing facility is the bottleneck under the exhaustive testing policy, and

$$r(M - \frac{\alpha b}{a-1}) - c_T > \frac{c_F}{(1-q)L}, \tag{11}$$

so that exhaustive testing is profitable.

3. The Optimal Fixed Sample Size Screening Policy

Since the optimal solution (λ, u) to problem (4)-(6) is difficult to obtain, we restrict ourselves in this section to a *fixed sample size* screening policy, which is denoted by $u^{n,B}$. Under this strategy, the number of wafers tested from a lot is $\min\{n, l\}$. If the total

number of defective chips found in these wafers is less than or equal to B, then the remaining wafers in the lot are tested; otherwise, the remaining wafers are discarded.

Standard calculations show that

$$f(s_{n-1}) = \frac{\Gamma(a+(n-1)\alpha)}{\Gamma(a)\Gamma((n-1)\alpha)} \frac{b^a s_{n-1}^{(n-1)\alpha-1}}{(b+s_{n-1})^{a+(n-1)\alpha}}, \quad s_{n-1} \geq 0, \tag{12}$$

is the probability density function for the number of bad chips on the first $n-1$ wafers tested, and

$$E[x_n|s_{n-1}] = \frac{\alpha(b+s_{n-1})}{a+(n-1)\alpha-1} \tag{13}$$

is the expected number of defective chips on the nth wafer, given that s_{n-1} defective chips are found on the first $n-1$ wafers. Also, the probability that a lot entering the testing facility has l wafers is given by

$$H(l) = \begin{pmatrix} L \\ l \end{pmatrix} l^{1-q}(L-l)^q. \tag{14}$$

Hence, the expected profit per lot of wafers is

$$\begin{aligned} V(u^{n,B}) &= \sum_{l=0}^{n} H(l)\{rl(M - E[x_n]) - lc_T\} \\ &+ \sum_{l=n+1}^{L} H(l)r\{n(M - E[x_n]) + (l-n)\int_{s_n \leq B}(M - E[x_{n+1}|s_n])f(s_n)ds_n\} \\ &- \sum_{l=n+1}^{L} H(l)c_T\{n + (l-n)\int_{s_n \leq B} f(s_n)ds_n\} \end{aligned} \tag{15}$$

and the expected number of wafers tested per lot is

$$N(U^{n,B}) = \sum_{l=0}^{n} H(l)l + \sum_{l=n+1}^{L} H(l)\{n + (l-n)\int_{s_n \leq B} f(s_n)ds_n\}. \tag{16}$$

Thus, problem (4)-(6) reduces to

$$\max_{\lambda,n,B} \quad \lambda(V(u^{n,B}) - c_F) \tag{17}$$

$$\text{subject to} \quad \lambda \leq \mu_F \tag{18}$$

$$\lambda N(u^{n,B}) \leq \mu_T, \tag{19}$$

which is equivalent to

$$\max_{n,B} \min\{\mu_F, \mu_T/N(u^{n,B})\}(V(u^{n,B}) - c_F). \tag{20}$$

Since a closed form solution to (20) appears to be unattainable, we exhaustively enumerate over the integer values $\{n, B : 0 \leq n \leq L; 0 \leq B \leq nM\}$ to find the optimal solution. The calculations are considerably streamlined by observing that

$$\begin{aligned} V(u^{n,B}) &= V(u^{n,B-1}) + \sum_{l=n+1}^{L} H(l)r(l-n) \int_{B-1}^{B} (M - E[x_{n+1}|s_n]) f(s_n) ds_n \\ &- \sum_{l=n+1}^{L} H(l)c_T(l-n) \int_{B-1}^{B} f(s_n) ds_n \end{aligned} \tag{21}$$

and

$$N(u^{n,B}) = N(u^{n,B-1}) + \sum_{l=n+1}^{L} H(l)(l-n) \int_{B-1}^{B} f(s_n) ds_n. \tag{22}$$

Hence, for each B, only $\int_{B-1}^{B} E[x_{n+1}|s_n] f(s_n) ds_n$ and $\int_{B-1}^{B} f(s_n) ds_n$ have to be calculated.

4. The Optimal Solution

In this section, a computational procedure is developed to solve problem (4)-(6), which is essentially an optimal stopping problem embedded within a mathematical program. First we reformulate the problem into the equivalent two-step maximization problem

$$\max_{0 \leq \lambda \leq \mu_F} \lambda(V_\lambda - c_F) \tag{23}$$

where

$$V_\lambda = \max_u V(u) \tag{24}$$

$$\text{subject to } N(u) \leq \frac{\mu_T}{\lambda}. \tag{25}$$

Proposition 1. *(λ^*, u^*) is an optimal solution to problem problem (4)-(6) if and only if u^* is an optimal solution to problem (24)-(25) with $\lambda = \lambda^*$ in (25), and λ^* is*

the optimal solution to (23). The optimal objective function value is the same for both problems.

Proof. If (λ^*, u^*) is an optimal solution to problem (4)-(6), then u^* satisfies (25) with $\lambda = \lambda^*$ and, for any screening policy u satisfying this condition, we have

$$\lambda^*(V(u^*) - c_F) \geq \lambda^*(V(u) - c_F) \tag{26}$$

or

$$V(u^*) \geq V(u). \tag{27}$$

Hence, u^* is an optimal solution to problem (24)-(25) with $\lambda = \lambda^*$ in (25), and $V(u^*) = V_{\lambda^*}$.

Observe that

$$\lambda^*(V(u^*) - c_F) \geq \lambda(V(u) - c_F) \tag{28}$$

for all λ and u satisfying (4) and (5). Fixing λ and maximizing over u subject to (25) yields

$$\lambda^*(V_{\lambda^*} - c_F) \geq \lambda(V_\lambda - c_F) \tag{29}$$

for all λ such that $0 \leq \lambda \leq \mu_F$. Therefore, λ^* is an optimal solution to (23) and the optimal objective function value is the same for both problems.

Conversely, if u^* is an optimal solution to (24)-(25) and λ^* is an optimal solution to (23), then they jointly satisfy constraints (4) and (5). For any other feasible solution (λ, u) to (4)-(6), we have

$$\lambda(V(u) - c_F) \leq \lambda(V_\lambda - c_F) \leq \lambda^*(V_{\lambda^*} - c_F) = \lambda^*(V(u^*) - c_F), \tag{30}$$

which implies that (λ^*,u^*) is an optimal solution to (4)-(6), and the optimal objective function value is the same for both problems. ∎

Let u^0 be the screening policy that maximizes the function $V(u)$ defined in (2). Maximizing $V(u)$ is an optimal stopping problem and will be discussed later in this section.

Proposition 2. *If $N(u^0) \leq \mu_T/\mu_F$, then the optimal solution to (23)-(25) is $u^* = u^0$ and $\lambda^* = \mu_F$.*

Proof. Since the screening strategy u^0 maximizes $V(u)$ with no side constraints, it also maximizes (24)-(25) for all $\lambda \in [0, \mu_T/N(u^0)]$. Since $N(u^0) \leq \mu_T/\mu_F$, it follows that u^0 optimizes (24)-(25) for all $\lambda \in [0, \mu_F]$. By (11), the exhaustive testing policy is profitable, and hence $V(u^0) > 0$ and setting $\lambda = \mu_F$ optimizes (23). ∎

Thus, when $N(u^0) \leq \mu_T/\mu_F$, the probing facility is not used to its full effective capacity, and the solution to (4)-(6) is obtained by solving a single optimal stopping problem. We now consider the more interesting situation where $N(u^0) > \mu_T/\mu_F$. Since u^0 optimizes (24)-(25) for all $\lambda \in [0, \mu_T/N(u^0)]$, (23) can be replaced by

$$\max_{\mu_T/N(u^0) \leq \lambda \leq \mu_F} \lambda(V_\lambda - c_F). \tag{31}$$

If problem (24)-(25) can be solved efficiently for a given λ, then a one-dimensional search over $\lambda \in [\mu_T/N(u^0), \mu_F]$ for the largest value of $\lambda(V_\lambda - c_F)$ will yield an optimal solution to our original problem. Since (24)-(25) is a constrained optimal stopping problem, we solve this problem by employing a Lagrangian approach. Let γ be the Lagrange multiplier for constraint (25) and define

$$g^\gamma(x_n, u_n) = \begin{cases} 0 & \text{if } u_n = 0, \\ r(M - x_n) - c_T - \gamma & \text{if } u_n = 1. \end{cases} \tag{32}$$

Notice that γ plays the role of an additional testing cost, so that the total testing cost per wafer is $c_T + \gamma$. Define the Lagrangian function

$$\begin{aligned} V^\gamma(u) &= V(u) - \gamma N(u) \\ &= E[\sum_{n=1}^{L} g(x_n, u_n)] - \gamma N(u) \\ &= E[\sum_{n=1}^{L} g^\gamma(x_n, u_n)], \end{aligned} \tag{33}$$

and consider the *Lagrangian problem*

$$\max_u V^\gamma(u). \tag{34}$$

Proposition 3. *If the screening policy $u^*(\gamma)$ solves the Lagrangian problem for some $\gamma \geq 0$, and*

$$N(u^*(\gamma)) = \frac{\mu_T}{\lambda}, \tag{35}$$

then $u^*(\gamma)$ is the optimal solution to problem (24)-(25).

Proof. For any screening strategy u satisfying (25),

$$\begin{aligned} V(u) &\leq V(u) - \gamma N(u) + \gamma \frac{\mu_T}{\lambda} \\ &\leq V(u^*(\gamma)) - \gamma N(u^*(\gamma)) + \gamma \frac{\mu_T}{\lambda} \\ &= V(u^*(\gamma)). \ \blacksquare \end{aligned} \tag{36}$$

Since γ enters the Lagrangian problem as an additional testing cost, it is not hard to show that the optimal objective function value in (34) is a continuous nonincreasing function of γ. The proof of the following proposition relies on this fact and the conjecture that the optimal expected number of wafers tested per lot $N(u^*(\gamma))$ is also a continuous, nonincreasing function of γ. Although this conjecture has been borne out in our numerical study and seems as intuitively obvious as the continuity and monotonicity of the optimal objective function value, the awkward expression for $N(u^*(\gamma))$ in (52) has prevented us from providing a rigorous proof.

Proposition 4. *If $N(u^0) > \mu_T/\mu_F$, then there exists a $\bar{\gamma} \in (0, rM]$ such that $u^*(\bar{\gamma})$ is an optimal solution to (24)-(25) with start rate $\lambda = \mu_F$.*

Proof. As γ increases from 0 to rM, $V^\gamma(u^*(\gamma))$ decreases from $V(u^0)$ to 0, since the optimal solution to (34) is to discard all wafers when $\gamma = rM$. Similarly, if our conjecture is correct, $N(u^*(\gamma))$ decreases from $N(u^0)$ to 0 as γ increases from 0 to rM. Since $0 < \mu_T/\mu_F < N(u^0)$, there must be a $\bar{\gamma} \in (0, rM]$ for which $N(u^*(\bar{\gamma})) = \mu_T/\mu_F$. By Proposition 3, $u^*(\bar{\gamma})$ is an optimal solution to (24)-(25) with start rate $\lambda = \mu_F$. $\blacksquare$

Propositions 3 and Proposition 4 can be combined to develop a search procedure for solving problem (4)-(6). For fixed $\lambda \in [\mu_T/N(u^0), \mu_F]$, we solve (34) and search for that γ for which $N(u^*(\gamma)) = \mu_T/\lambda$. Proposition 4 guarantees the existence of such a

γ in the interval $[0, \bar{\gamma}]$. Then, we evaluate the objective function in (31) and search for a λ^* that has the largest objective function value. However, the search over λ can be accomplished *simultaneously* as we search over γ. For each $\gamma \in [0, \bar{\gamma}]$, we solve (34) and let $\lambda = \mu_T / N(u^*(\gamma))$. By Proposition 3, $u^*(\gamma)$ is the optimal solution to (24)-(25) with this value of λ, and the objective function in (31) can be evaluated. Since for every $\lambda \in [\mu_T / N(u^0), \mu_F]$ there exists a γ in the interval $[0, \bar{\gamma}]$ such that $N(u^*(\gamma)) = \mu_T / \lambda$, every $\lambda \in [\mu_T / (1-q)L, \mu_F]$ is searched as all $\gamma \in [0, \bar{\gamma}]$ are searched. Thus, one search over $\gamma \in [0, \bar{\gamma}]$ is sufficient to find the optimal start rate λ^* to (31) and the optimal screening policy u^* to (24)-(25). Readers can find an outline of this algorithm at the end of this section.

We now focus on solving the Lagrangian problem (34). Let

$$G(l) = 1 - \sum_{1}^{l} H(k) \tag{37}$$

be the probability that a lot has more than l wafers, and let

$$f(x_n | s_{n-1}) = \frac{\Gamma(a + n\alpha)}{\Gamma(\alpha)\Gamma(a + (n-1)\alpha)} \frac{(b + s_{n-1})^{a+(n-1)\alpha} (x_n)^{\alpha - 1}}{(b + s_{n-1} + x_n)^{a+n\alpha}}, \quad x_n \geq 0, \tag{38}$$

denote the posterior probability density for the number of bad chips on the nth wafer, given that s_{n-1} bad chips are found on the first $n-1$ wafers. If $V_n^\gamma(s_n)$ represents the expected profit obtained from wafers $n+1, ..., L$, given that s_n bad chips were detected on the first n wafers, then $u^*(\gamma)$ and $V^\gamma(u^*(\gamma))$ can be found by solving the dynamic programming equations

$$V_L^\gamma(s_L) = 0 \tag{39}$$

$$V_n^\gamma(s_n) = \max\{0, G(n) \int_0^\infty [r(M - x_{n+1}) - c_T - \gamma + V_{n+1}^\gamma(s_n + x_{n+1})] f(x_{n+1} | s_n) dx_{n+1}\},$$
$$n = L-1, \ldots, 1, \text{ and} \tag{40}$$

$$V^\gamma(u) = V_0^\gamma(0) = \max\{0, G(0) \int_0^\infty [r(M - x_1) - c_T - \gamma + V_1^\gamma(x_1)] f(x_1) dx_1\}. \tag{41}$$

After n wafers have been tested, we can discard the remaining wafers and obtain no profit. If the lot has more than n wafers, then we can continue testing; if wafer $n+1$ contains x_{n+1} bad chips, then the immediate profit is $r(M - x_{n+1}) - c_T - \gamma$ and the expected

future profit is $V^{\gamma}_{n+1}(s_n + x_{n+1})$. These equations also reveal structural properties of the optimal solution to (34), which are discussed in the two propositions below.

Proposition 5. *The optimal policy* $u^*(\gamma) = (u^*_1(\gamma), ..., u^*_L(\gamma))$ *is*

$$u^*_n = 1 \qquad \text{if } s_{n-1} \leq B^{\gamma}_{n-1}. \tag{42}$$

$$u^*_n = 0 \qquad \text{if } s_{n-1} > B^{\gamma}_{n-1}, \tag{43}$$

where the stopping boundary $B^{\gamma}_n \geq -1, n = 1, ..., L$, *and* $B^{\gamma}_{n-1} = -1$ *indicates that wafer* n *is not tested under any circumstances.*

Proof. We only need to show that $V^{\gamma}_n(s_n)$ is nonincreasing in s_n, which is done by a backward induction on n. It is trivially true for $n = L$. Suppose it is true for $n + 1$, and consider the difference $V^{\gamma}_n(s_n + 1) - V^{\gamma}_n(s_n)$. In order to prove that this quantity is nonpositive, the following properties of the conditional density $f(x_{n+1}|s_n)$ are required. For $n = 1, \ldots, L - 1$ and $s_n \geq 0$, there exists $\bar{x}_{n+1} > 1$ such that

$$f(x_{n+1} - 1|s_n + 1) \geq f(x_{n+1}|s_n), \text{ for } x_{n+1} \geq \bar{x}_{n+1}, \text{ and} \tag{44}$$

$$f(x_{n+1} - 1|s_n + 1) < f(x_{n+1}|s_n), \text{ for } x_{n+1} < \bar{x}_{n+1}. \tag{45}$$

These inequalities can be verified using (38). By (40), it suffices to consider the difference

$$\begin{aligned}
&\int_0^{\infty} [(r(M - x_{n+1}) - c_T - \gamma + V^{\gamma}_{n+1}(s_n + 1 + x_{n+1}) f(x_{n+1}|s_n + 1) dx_{n+1}] \\
&- \int_0^{\infty} [(r(M - x_{n+1}) - c_T - \gamma + V^{\gamma}_{n+1}(s_n + x_{n+1}) f(x_{n+1}|s_n) dx_{n+1}] \\
&= -\frac{r\alpha}{a + n\alpha - 1} + \int_0^{\infty} V^{\gamma}_{n+1}(s_n + 1 + x_{n+1}) f(x_{n+1}|s_n + 1) dx_{n+1} \\
&- \int_0^{\infty} V^{\gamma}_{n+1}(s_n + x_{n+1}) f(x_{n+1}|s_n) dx_{n+1} \\
&< \int_0^{\infty} [V^{\gamma}_{n+1}(s_n + 1 + x_{n+1}) - V^{\gamma}_{n+1}(s_n + \bar{x}_{n+1})] f(x_{n+1}|s_n + 1) dx_{n+1} \\
&- \int_0^{\infty} [V^{\gamma}_{n+1}(s_n + x_{n+1}) - V^{\gamma}_{n+1}(s_n + \bar{x}_{n+1})] f(x_{n+1}|s_n) dx_{n+1}.
\end{aligned} \tag{46}$$

Changing the integration variable in the first integral from x_{n+1} to $x_{n+1}+1$ and combining

the two integrals in (46), we get

$$\int_1^{\infty} [V_{n+1}^{\gamma}(s_n + x_{n+1}) - V_{n+1}^{\gamma}(s_n + \bar{x}_{n+1})][f(x_{n+1} - 1|s_n + 1) - f(x_{n+1}|s_n)]dx_{n+1}$$
$$- \int_0^1 [V_{n+1}^{\gamma}(s_n + x_{n+1}) - V_{n+1}^{\gamma}(s_n + \bar{x}_{n+1})]f(x_{n+1}|s_n)dx_{n+1}. \quad (47)$$

The two terms inside the first integral have opposite signs by (44)-(45) and the induction hypothesis; hence, the first integral is nonpositive. The second integral is nonnegative because, by the induction assumption, $V_{n+1}^{\gamma}(s_n + x_{n+1}) \geq V_{n+1}^{\gamma}(s_n + \bar{x}_{n+1})$ for $0 \leq x_{n+1} \leq 1 < \bar{x}_{n+1}$. Therefore, (47) is nonpositive, and the induction is verified. ∎

The following proposition establishes monotonicity of the optimal stopping boundary, and is used to streamline the dynamic programming algorithm. The proof is similar to the proof of Proposition 5, and is omitted.

Proposition 6. *The optimal stopping boundary satisfies*

$$B_0^{\gamma} \leq B_1^{\gamma} \leq \ldots \leq B_{L-1}^{\gamma}. \quad (48)$$

The dynamic programming equations (39)-(41) involve L functions of continuous variables. In the numerical computations, we discretize the continuous variables and approximate the integrals by finite summations. Two observations are helpful in reducing the amount of computation. First, the final boundary point can be explicitly derived, and equals

$$B_{L-1}^{\gamma} = (a + (L-1)\alpha - 1)(M - (\gamma + c_T)/r)/\alpha - b. \quad (49)$$

Also, since $V_n(s_n)$ is nonincreasing in s_n, we calculate $V_n^{\gamma}(s_n)$ starting from $s_n = 0$, and if s_n is found such that $V_n^{\gamma}(s_n) = 0$, then we set $V_n(x) = 0$ for all $x \in (sn, nM]$.

After the optimal solution $u^*(\gamma)$ to the Lagrangian problem is derived, we need to determine $N(u^*(\gamma))$, which is the expected number of wafers that are tested per lot. Notice that the optimal boundary point $B_0^{\gamma} = -1$ or 0. If $B_0^{\gamma} = -1$, then the screening policy $u^*(\gamma)$ cannot be optimal for the original problem (4)-(6), by (11). If $B_0^{\gamma} = 0$, then

$$C_n = \int_{s_1=0}^{B_1} \ldots \int_{s_n=0}^{B_n} f(s_n - s_{n-1}|s_{n-1}) \ldots f(s_1)ds_n \ldots ds_1, \; n = 1, \ldots, L-1, \quad (50)$$

and the probability that testing ceases after the nth wafer is

$$T_n = C_{n-1} - C_n, \; n = 1, \ldots, L-1, \tag{51}$$

where $C_0 = 1$. Then the expected number of wafers tested per lot is

$$N(u^*(\gamma)) = \sum_{n=1}^{L-1} nT_n + L(1 - \sum_{n=1}^{L-1} T_n). \tag{52}$$

We conclude this section with an outline of the algorithm that solves the original optimization problem (4)-(6).

Algorithm:

Step 1. Let $\gamma = 0$ and set $\gamma^* = 0$.

Step 2. Find the optimal solution $u^*(\gamma)$ to the Lagrangian problem (34) and the optimal objective function value $V^\gamma(u^*(\gamma))$. If the optimal boundary $B_0^\gamma = -1$, then stop. The optimal start rate is λ^{γ^*} and the optimal screening policy is $u^*(\gamma^*)$.

Step 3. Calculate $N(u^*(\gamma))$ using (52), and define $\lambda^\gamma = \mu_T / N(u^*(\gamma))$.

Step 4. Compute the objective function value for the original problem,

$$P_\gamma = \lambda^\gamma [V^\gamma(u^*(\gamma)) + \gamma N(u^*(\gamma)) - c_F]. \tag{53}$$

If P_γ is the maximum over all P_γ's calculated thus far, then let $\gamma^* = \gamma$. Change γ to $\gamma + \delta$, where δ is a small step variable, and go to step 2.

Notice that the algorithm is guaranteed to terminate, since $B_0^\gamma = -1$ when $\gamma = rM$. If we could prove that P_γ is concave in γ, then a binary search, rather than an exhaustive search, over $\gamma \in [0, rM]$ could be employed, thereby saving a considerable amount of computation. For all five data sets considered in the next section, P_γ is indeed concave with respect to γ. Although we have been able to prove that the optimal value function $V^\gamma(u^*(\gamma))$ is decreasing and concave in γ, we have been unable to prove the concavity of P_γ; our obstacle is again the expression for $N(u^*(\gamma))$ in (52).

5. Numerical Results

In this section, we test the optimal fixed sample size screening strategy and the optimal sequential screening strategy on five sets of yield data, where each set has about 10 lots and each lot has less than 25 wafers. The data sets, denoted by C1,C1.5,C2,C2.5 and C3, were obtained by Bohn from the same factory producing the same product in five different time periods. For each of the five data sets, maximum likelihood estimation is used to obtain values of the gamma parameters α, a and b. More specifically, the following procedure is followed for each data set. If a data set contains m lots, then the maximum likelihood estimates $(\alpha_1, ..., \alpha_m)$ and $(\beta_1, ..., \beta_m)$ are obtained from the number of defective chips on each wafer in the set. We estimate α by $\hat{\alpha}$, which is the median of $(\alpha_1, ..., \alpha_m)$, and then recompute the estimates $(\beta_1, ..., \beta_m)$ by assuming that the number of defective chips on each wafer in lot k is a gamma random variable with known shape parameter $\hat{\alpha}$ and scale parameter β_k. Finally, the revised estimates $(\beta_1, ..., \beta_m)$ are used to obtain maximum likelihood parameter estimates $\hat{a}$ and $\hat{b}$. These parameter estimates $\hat{\alpha}, \hat{a}$ and $\hat{b}$ are not reported here for reasons of confidentiality. We also performed the identical estimation procedure, but chose $\hat{\alpha}$ to be the mean, rather than the median, of $(\alpha_1, ..., \alpha_m)$; the profitability results for this case were quite similar to the results obtained from the original procedure and are omitted.

As mentioned earlier, we assume that when the the probing facility is working at its effective capacity under the exhaustive testing policy, the fab is working at 90% of its effective capacity. The wafer scrap rate is 5%, the variable probing cost per wafer is 3% of the variable fabrication cost per wafer and the revenue from a wafer containing all good chips is 10 times the wafer's variable production cost. These parameter values are based on discussions with a variety of semiconductor managers and engineers, and are used to derive the optimal fixed sample size strategy and the optimal sequential screening strategy for each of the five data sets. Both screening policies vary little over the five data sets, and Figure 3 illustrates the two policies for data set C1. The stopping boundary characterizing the optimal sequential screening policy is nearly linear, but slightly convex, for every data

set, and the slope increases with increased lot-to-lot variation. Since the slope determines the acceptable yield level, as the lot-to-lot variability increases, more testing is required before discarding the remaining wafers in the lot. The average acceptable yield level over the five data sets is 13.7% lower than the overall average yield. The optimal fixed sample size policy samples either or 6 wafers from every lot in each data set, and requires a slightly higher yield level to continue testing than the optimal sequential policy. Since the fixed sample size policy stops monitoring yield after a lot is considered acceptable while the sequential screening strategy monitors yield continuously, it is not surprising that the former only accepts lots of expected higher yield level.

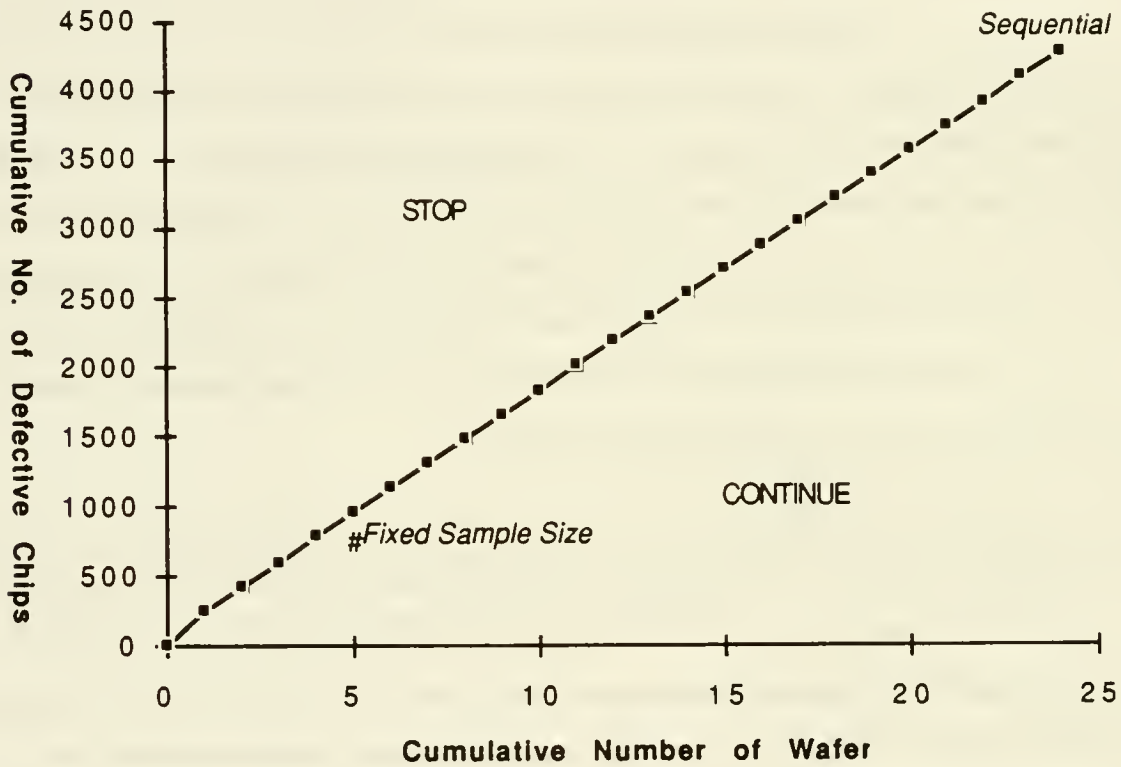

Figure 3. Optimal screening policies for data set C1.

Recall that the strategy commonly used in industry is to perform exhaustive testing and to choose the start rate so that the testing facility works at its effective capacity.

Before reporting our profit results, it is useful to determine an upper bound on the profit increase that can be achieved relative to this straw strategy. Let $\lambda^E = \mu_T/(1-q)L$ denote the arrival rate under the exhaustive testing strategy, and let

$$y = \frac{M - \frac{\alpha b}{a-1}}{M} \tag{54}$$

denote the average incoming yield. Then an upper bound on the relative profit increase for any strategy is

$$\begin{aligned}
\frac{\lambda^*(V(u^*) - c_F) - \lambda_E(V(u^E) - c_F)}{\lambda^E(V(u^E) - c_F)} &= \frac{\lambda^* - \lambda^E}{\lambda^E} + \frac{\lambda^*}{\lambda^E}\left(\frac{V(u^*) - V(u^E)}{V(u^E) - c_F}\right) \\
&\leq \frac{\mu_F - \lambda^E}{\lambda^E} + \frac{\mu_F}{\lambda^E}\left(\frac{V(u^*) - V(u^E)}{rL(1-q)My - Lc_T - c_F}\right) \\
&\leq \frac{\mu_F - \lambda^E}{\lambda^E} + \frac{\mu_F}{\lambda^E}\left(\frac{Lc_T(1-y)}{rL(1-q)My - Lc_T - c_F}\right). \qquad (55)
\end{aligned}$$

Since we assumed that $Lc_T = 0.03c_F$, $rLM = 10c_F$, $q = 0.05$ and the fab is at 90% of its effective capacity under the exhaustive strategy, the upper bound equals

$$\Delta P_{\max} = \frac{1}{9} + \frac{10}{9}\left(\frac{0.03(1-y)}{9.5y - 1.03}\right). \tag{56}$$

This quantity increases from 1/9 when yield is 100% to ∞ as yield approaches the critical level of 10.58% required for profitability in (11). Although the exact average yield cannot be revealed, the yield was greater than 36% for all five data sets, and hence $\Delta P_{\max}$ is between 11.1% and 12.0%.

Under the heading "Theoretical Calculations", Table I reports the profit increases relative to the exhaustive testing strategy obtained by the two proposed strategies for all five data sets. We also display $\rho_F = \lambda/\mu_F$ and $\rho_T = (\lambda N(u)/\mu_T)$, which represent the *effective capacity utilization* of the fab and testing facility, respectively. It can be seen that for every data set, both facilities work at their effective capacity. However, the profit increases are rather small: out of a potential 11.1% to 12.0% increase, only a 2% to 3% increase is achieved. Also, the difference in performance between the two proposed policies is relatively small; the fixed sample size policy averages a 2.24% profit increase over the five data sets, compared to 2.51% for the optimal strategy.

Table 1. Numerical results.

Data Set		Theoretical Calculations		Simulation Results	
		Fixed Sample Size	Sequential	Fixed Sample Size	Sequential
C1	ΔP	1.80	2.09	1.23	1.37
	ρ_F	1.000	1.000	1.000	1.000
	ρ_T	1.000	1.000	1.000	1.000
C1.5	ΔP	2.18	2.43	1.04	-4.02
	ρ_F	1.000	1.000	1.000	1.000
	ρ_T	1.000	1.000	1.000	0.948
C2	ΔP	2.99	3.27	0.70	-1.90
	ρ_F	1.000	1.000	1.000	1.000
	ρ_T	1.000	1.000	0.989	0.958
C2.5	ΔP	2.15	2.43	0.75	0.90
	ρ_F	1.000	1.000	1.000	1.000
	ρ_T	1.000	1.000	1.009	1.007
C3	ΔP	2.08	2.35	2.45	0.14
	ρ_F	1.000	1.000	1.000	1.000
	ρ_T	1.000	1.000	0.999	0.971

ΔP: percentage profit increase over exhaustive testing strategy

ρ_F: utilization of effective capacity of the fab

ρ_T: utilization of effective of the testing facility

Since we do not know the order in which the wafers in each lot of the five data sets were actually tested, it is difficult to test our proposed strategies directly on the actual data. Therefore, we reverted to simulation, and generated 100 lots from each sample lot by randomly shuffling the wafers in the lot. For each data set and each screening policy, the average number of good chips obtained per lot and the average number of

wafers tested per lot were recorded. These quantities and the theoretically calculated start rates were then used to calculate the profit increases that are reported in Table I under the heading "Simulation Results".

When the derived policies are tested on the actual data, two undesirable things can occur. If the yield model underestimates the number of discarded wafers, then the testing facility is underutilized and a feasible, but suboptimal, strategy is obtained. If the yield model overestimates the number of discarded wafers, then the testing facility is overutilized, and an infeasible strategy can result. Referring to Table I, we see that the yield model correctly predicts the average number of discarded wafers per lot under the fixed sample size policy for three of the five data sets, and is off by about 1% on data sets C2 and C2.5. However, the yield model is less accurate under the sequential policy, underestimating the average number of discarded wafers per lot by 3-5% in three of the five data sets. In these cases, the resulting profit is sometimes less than under exhaustive testing. Both policies overestimated the number of discarded wafers in data set C2.5 and the resulting strategy is not feasible; hence, the profit increases reported for this data set correspond to a reduced start rate that maintains feasibility. That is, the profit increase of 0.75 (0.90, respectively) was achieved by reducing the start rate so that $\rho_F = 0.991$ (0.993, respectively) and $\rho_T = 1.000$. The average profit increase over the five data sets for the fixed sample size strategy is 1.23% in the simulation study, about 1% below the corresponding improvement achieved in the theoretical calculations. The sequential strategy averages a 0.70% profit decrease relative to the straw strategy, because of the underutilization of the testing facility in cases C1.5 and C2. Hence, in addition to being easier to derive and to implement than the sequential strategy, the fixed sample size strategy performs nearly as well in the analytical calculations, and appears to be more robust in our limited simulation study.

As a point of reference, we also considered the beta-binomial yield model, where the number of bad chips on each wafer is modeled as a binomial random variable. This yield model significantly overestimated the average number of wafers tested per lot: the

average value over the five data sets of ρ_T under the sequential strategy in the simulation study was only 0.861, which led to an average profit decrease of 11.68%.

Acknowledgment

We are deeply indebted to Roger Bohn, who generously gave us his yield data. This research is supported by a grant from the Leaders for Manufacturing Program at MIT and National Science Foundation grant DDM-9057297.

References

Albin, S. and D. J. Friedman. 1989. The Impact of Clustered Defect Distribution in IC Fabrication. *Management Science* **35**, 1066-1078.

Atherton, R. W. and J. E. Dayhoff. 1985. Introduction to Fab Graph Structures. *ECS Abstracts.*

Bohn, R. E. 1991. Noise and Learning in Semiconductor Manufacturing. Center for Technology Policy and Industrial Development, MIT, Cambridge, MA.

Chen, H, M., J. Harrison, A. Mandelbaum, A. van Ackere, and L. M. Wein. 1988. Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication. *Operations Research* **36**, 202-215.

Cunningham, J. A. 1990. The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing. *IEEE Trans. on Semiconductor Manufacturing* **3**, 60-72.

Ferris-Prabhu, A. V., L. D. Smith, H. A. Bonges, and J. K. Paulsen. 1987. Radial Yield Variations in Semiconductor Wafers. *IEEE Circuits and Devices Magazine*, March, 42-47.

Glassey, R. C. and M. G. C. Resende. 1988. Closed-Loop Release Control for VLSI Circuit Manufacturing. *IEEE Trans. on Semiconductor Manufacturing* **1**, 36-46.

Goldratt, E. M. and J. Cox. 1984. *The Goal.* North River Press, Inc., Croton-on-Hudson, NY.

Longtin, M., L. M. Wein, and R. E. Welsch. 1992. Sequential Screening in Semiconductor Manufacturing, II: Exploiting Spatial Dependence. Sloan School of Management, MIT, Cambridge, MA.

Mallory, C. L., D. S. Perloff, T. F. Hasan, and R. N. Stanley. 1983. Spatial Yield Analysis in Integrated Circuit Manufacturing. *Solid State Technology*, November, 121-127.

Osburn, C. M., H. Berger, R. P. Donovan, and G. W. Jones. 1988. The Effects of Contamination on Semiconductor Manufacturing Yield. *The Journal of Environmental Sciences*, March/April, 45-57.

Wein, L. M. 1988. Scheduling Semiconductor Wafer Fabrication. *IEEE Trans. on Semiconductor Manufacturing* **1**, 115-130.